12-2001

Nearest Neighbor Methods

Philip M. Dixon
Iowa State University, pdixon@iastate.edu

Nearest Neighbor Methods

Abstract
Nearest neighbor methods are a diverse group of statistical methods united by the idea that the similarity between a point and its nearest neighbor can be used for statistical inference. This review article summarizes two common environmetric applications: nearest neighbor methods for spatial point processes and nearest neighbor designs and analyses for field experiments. In spatial point processes, the appropriate similarity is the distance between a point and its nearest neighbor. Given a realization of a spatial point process, the mean nearest neighbor distance or the distribution of distances can be used for inference about the spatial process. One common application is to test whether the process is a homogeneous Poisson process. These methods can be extended to describe relationships between two or more spatial point processes. These methods are illustrated using data on the locations of trees in a swamp hardwood forest. In field experiments, neighboring plots often have similar characteristics before treatments are imposed. This similarity can be estimated and used to remove bias and increase the precision of treatment comparisons. Some of the commonly used nearest neighbor methods are briefly described.

Disciplines
Statistics and Probability

Comments
This preprint was published as Philip M. Dixon, "Nearest Neighbor Methods", *Encyclopedia of Environmetrics* (2002): :1370-1383.

Nearest Neighbor Methods

Philip M. Dixon

Department of Statistics

Iowa State University

20 December 2001

"Nearest neighbor methods" include at least six different groups of statistical methods. All have in common the idea that some aspect of the similarity between a point and its nearest neighbor can be used to make useful inferences. In some cases, the similarity is the distance between the point and its nearest neighbor; in others, the appropriate similarity is based on other identifying characteristics of the points. I will discuss in detail nearest neighbor methods for spatial point processes and field experiments because these are commonly used in biology and environmetrics. I will very briefly discuss nearest-neighbor designs for field experiments, in which each pair of treatments occurs as neighbors equally frequently. I will not discuss nearest neighbor estimates of probability density functions [23], nearest neighbor methods for discrimination or classification [59, pp. 191-201], or nearest neighbor linkage (i.e. simple linkage) in hierarchical clustering [32, , pp. 57-60]. Although these last three methods have been applied to environmetric data, they are much more general.

Nearest neighbor methods for spatial point processes

Spatial point process data describe the locations of "interesting" events, and (possibly) some information about each event. Some examples include locations of tree trunks (e.g. [52]), locations of bird nests (e.g. [11]), locations of pottery shards, and locations of cancer cases (e.g. [20]). I will focus on the most common case where the location is recorded in two dimensions (x, y). Similar techniques can be used for three dimensional data (e.g. locations of galaxies in space) or one dimensional data (e.g. nesting sites along a coastline or along a riverbank). Usually, the locations of all events in a defined area are observed (completely mapped data), but occasionally only a subset of locations are observed (sparsely sampled data). Univariate point process data includes only the locations of the events; marked point process data includes additional information about the event at each location [65]. For example, the species may be recorded for each tree, some cultural indentification may be recorded for each pottery shard, and nest success or nest failure may be recorded for each bird nest.

Location or marked location data can be used to answer many different sorts of questions. The scientific context for a question depends on the area of application, but the questions can be grouped into general categories. One very common category of questions concerns the spatial pattern of the observations. Are the locations spatially clustered? Do they tend to be regularly distributed, or are they random (i.e a realization of a homogeneous Poisson process)? A second common set of questions concerns the relationships between different types of events in a marked point process. Do two different species of tree tend to occur together? Are locations of cancer cases more clustered than a random subset of a control group? A third set of questions deals with the density (number of events per unit area). What is the average density of trees in an area? What does a map of density look like? Methods to answer each of these types of questions are discussed in the following sections.

Theoretical treatments of nearest neighbor methods for spatial point patterns can be found in [25, 18, 65]. Applications of nearest neighbor methods can be found in many papers and books, including [53, 64, 39, 66].

Describing and testing spatial patterns using completely mapped data.

Describing and testing spatial patterns of locations has a long history. Historically, the primary concern was with the question of randomness [1, 2, 15, 48]. Are locations randomly distributed throughout the study area (i.e. are the locations a realization of a Poisson process with homogeneous intensity)? Or, did do the locations indicate some structure (i.e. clustering or repulsion between locations). Because of the many connotations of randomness and the importance of a homogeneous Poisson process as a benchmark, it is commonly called 'Complete spatial randomness' or CSR.

In this section, I will describe nearest-neighbor tests based on completely mapped data. Locations of all events are recorded in an arbitrary study region. Often the study region is square or rectangular, but this is not a requirement. Tests for the less common case of sampled data are described in the next section.

Tests based on mean nearest neighbor distance

The distances between nearest neighbors provide information about the pattern of points. Define W as the distance from a randomly chosen event to the nearest other event in a homogenous Poisson process with intensity (expected # of points per unit area) of ρ. The pdf and cdf of W are

$$g(w) = 2\rho\pi w \exp(-\rho\pi w^2), \quad (1)$$

$$G(w) = 1 - e^{-\rho\pi w^2} \quad (2)$$

so the mean and the variance of W are

$$\mathrm{E}\ W = 1/(2\sqrt{\rho}) \tag{3}$$

and

$$\mathrm{Var}\ W = (4-\pi)/(4\pi\rho). \tag{4}$$

Based on these moments, Clark and Evans (1954) proposed a test of CSR. The conditional moments, E $W \mid \hat{\rho}$ and Var $W \mid \hat{\rho}$ are calculated by substituting the observed density, $\hat{\rho}$ = # total number of points / total area of study region, into (3) and (4). The observed mean nearest-neighbor distance, $\overline{w}$ is computed by identifying the nearest neighbor of each point, finding the distance between nearest neighbors, then averaging. Clark and Evans (1954) proposed that the standardized mean

$$Z_{CE} = \frac{\overline{w} - \mathrm{E}\ W \mid \hat{\rho}}{\sqrt{N^{-1}\mathrm{Var}\ W \mid \hat{\rho}}} \tag{5}$$

has a standard normal distribution if the process is CSR.

The Z_{CE} statistic and the many users of it ignore two problems: non-independence of some nearest-neighbor distances and edge effects. In a completely mapped area, many of the distances between nearest neighbors are correlated. The problem is most severe with reflexive nearest neighbors. Two points, A and B, are reflexive nearest neighbors when B is the nearest neighbor of A and A is the nearest neighbor of B [16]. Other authors have called these isolated nearest neighbors [51] or mutual nearest neighbors [61]. When A and B are reflexive nearest neighbors, each point has the value of W, which inflates the variance of the mean nearest-neighbor distance. This problem is not restricted to a few points. When points are CSR in 2 dimensions, approximately 62.15% of the points are reflexive nearest neighbors [16].

Edge effects arise because the distribution of W (2) assumes an unbounded area, but the observed nn distances are calculated from points in a defined study area. When a point is near the edge of the study area, it is possible that the true nearest neighbor is a point just outside the study area, not a more distant point that happens to be in the study area. Edge effects lead to overestimation (positive bias) of the mean nn distance. Edge effects can be practically important; neglecting them can alter conclusions about the spatial pattern (e.g. [10]).

Edge effects may be minimized by including a buffer area that surrounds the primary study area [15]. Nearest neighbor distances are only calculated for points in the primary study area, but locations in the buffer area are available as potential nearest neighbors. With a sufficiently large buffer area, this approach can eliminate edge effects, but it is wasteful since an appropriately large buffer area may contain many locations.

Using simulations, Donnelly [31] derived edge-corrected approximations to E $W \mid \hat{\rho}$ and Var $\overline{w} \mid \hat{\rho}$ when the

study region is rectangular,

$$\mathrm{E}\ W \mid \hat{\rho} \approx 0.5(A/N)^{1/2} + 0.0514\, P/N + 0.041\, P/N^{3/2} \quad (6)$$

$$\mathrm{Var}\ \overline{w} \mid \hat{\rho} \approx 0.0703\, A/N^2 + 0.037\, PA^{1/2}/N^{5/2}, \quad (7)$$

where N is the observed number of points, A is the area of the study region, and P is the total perimeter of the study region. These approximations can be used to test CSR by substituting them into (5), then comparing the z-score to a standard normal distribution. This test has reasonable power to detect departures from CSR [57].

One difficulty with tests based on the mean nn distance is that the mean is just a single summary of the pattern. Two point patterns may have the same mean nn distance, but one is CSR and the other is not. One such pattern would have a few patches of clustered points and an appropriate number of widely scattered individuals. The points in the clusters have small nn distances, but the widely scattered individuals have large nn distances. With the appropriate mix of clustered and scattered points, the mean nn distance could be exactly that given by (3).

Distribution of nearest-neighbor distances

An alternative is to consider the entire distribution, $G(w)$ of nearest neighbor distributions [24]. CSR can be tested by comparing the observed distribution function of nn distances, $\hat{G}(w)$, to the theoretical cdf (2). A variety of test statistics have been suggested, including Kolmogorov-Smirnov type statistics: $\sup_w \mid \hat{G}(w) - G(w) \mid$, Cramer-von Mises type statistics: $\int_w (\hat{G}(w) - G(w))^2$, or Anderson-Darling type statistics: $\int_w (\hat{G}(w) - G(w))^2 / G(w)(1 - G(w))$. The Kolmogorov-Smirnov statistic seems to be the most commonly used. The usual critical values for the 1-sample Kolmogorov-Smirnov test are not appropriate here because of non-independence of nearest neighbor distances, especially for reflexive nearest neighbors, as discussed above. A Monte-Carlo test must be used [26].

The Monte-Carlo approach computes the α-level critical value or the p-value by simulation. The number of points, N, is fixed at the observed number. N random locations are simulated as a realization of CSR in the study area, and the Kolmogorov- Smirnov (or other) test statistic is computed. This is repeated R times to give R values from the sampling distribution of the test statistic under the null hypothesis of CSR. The one-sided α-level critical value is the $\alpha(R+1)$'th largest value from the sampling distribution. The p-value is computed as (1+# random values larger than the observed value) / (1 + R).

Edge effects complicate the estimation of $\hat{G}(w)$. One solution is to include a border strip, as discussed above, but this may ignore a considerable amount of information. A variety of edge-corrected estimators of $G(w)$ have

been proposed; four of them are summarized in [18, pp. 613-4 and 637-8]. Edge corrections reduce the bias in the estimator, but they increase the sampling variance [40].

Although edge-corrected estimators are needed if the observed distribution function is compared to the theoretical distribution function under CSR (equation 2), they may not be needed for a Monte-Carlo test of CSR. A more powerful test of CSR is to use a non-edge corrected estimator (equation 2) and compare the biased estimate of $G(w)$ to the biased mean, $\overline{G}^*(w)$, under CSR [40]. The biased mean $\overline{G}^*(w)$ and point-wise prediction intervals are computed by simulation. Values of $\hat{G}(w)$ above the simulated mean indicate clustering of points (an excess of short distances to nearest neighbors). Values of $\hat{G}(w)$ below the simulated mean indicate regularity (few to no points with short distances to nearest-neighbors).

Graphical diagnostics based on the empirical distribution of nearest neighbor distances provide additional information about the spatial process. The most common graphical diagnostic is a quantile-quantile plot of either $G(w)$ or $\overline{G}^*(w)$ on the X axis and $\hat{G}(w)$ on the Y axis. $G(w)$ would be used when the object is to evaluate the fit of the theoretical nn distribution; $\overline{G}^*(w)$ would be used when the theoretical nn distribution is unknown and the object is to evaluate the fit to a process that can be simulated.

Monte-Carlo simulation of the spatial process provides both an estimate of $\overline{G}^*(w)$ and the sampling distribution of $\hat{G}^*(w)$ conditional on a specific spatial process. Quantiles of the sampling distribution of $\hat{G}^*(w)$ can be calculated for interesting distances, w. If the spatial process is simulated many times (e.g. 199 or 999), the quantiles can be estimated relatively precisely. The 0.05 and 0.95 quantiles can be approximated by the minimum and maximum of R=19 simulations. The 0.01 and 0.99 quantiles can be approximated by the minimum and maximum of R=99 simulations.

The simulated mean and quantiles of $\hat{G}^*(w)$ are plotted against the observed $\hat{G}(w)$. If the observed curve of $\hat{G}(w)$ falls entirely between the lower and upper bounds, there is no evidence against the null hypothesis (e.g. that the locations are a realization of CSR). An excursion outside the bounds indicates a departure from CSR. If the observed $\hat{G}(w)$ falls below the lower bound at short distances, there are too few nearest neighbors at short distances, which is consistent with a regular pattern, or one where there is inhibition of nearby points. If the observed $\hat{G}(w)$ lies above the upper bound at short distances, there are too many nearest neighbors at short distances, which is consistent with a clustered process.

The Monte-Carlo approach is not limited to testing CSR. It can be used to evaluate fit of any process that can be simulated. For example, a set of locations might be compared to a Poisson cluster process or a Strauss process [31, 25].

Nearest-neighbor methods have been extended in a variety of ways. Tests can be based on other functions of the nearest neighbor distances (e.g. squared nn distances [12], or the smallest nn distance [62]), but such tests have not been widely used. Distances to second nearest-neighbor, or the third nearest-neighbor, or perhaps an even further neighbor have been suggested as a way to look at patterns at a larger scale. The set of cdf's of distance to the nearest-neighbor, distance to the second nearest-neighbor, $\cdots$, distance to n-th nearest-neighbor is related to Ripley's K(t) function, another commonly used method to analyze spatial patterns [65, p 267]. Finally, the distance between a randomly chosen point and the nearest event also provides information about the spatial pattern [25, 66].

point-event distances

The point-event distribution, $F(x)$, considers the distance between a randomly chosen location (not the location of an event) and the nearest event. This can be estimated by choosing m locations in the study area and computing the distance from each location to its nearest neighbor. As with $G(w)$, edge effects complicate estimation of the cumulative distribution function. An edge-corrected estimator is

$$\hat{F}_R(x) = \#(x_I \leq x, d_I > x)/\#(d_I > x), \tag{8}$$

where x_I is the distance between a point and its neighboring event and d_I is the distance between a point and its nearest boundary. When the events are a realization of CSR, X, the point-event distance, and W, the nearest neighbor distance have the same distribution, so $F(x) = 1 - \exp(-\rho\pi x^2)$. However, the effects of deviations from randomness on $F(x)$ are opposite of those on $G(w)$. Values of $\hat{F}(x)$ above the expected value (8) indicate regularity. Values below the expected value indicate clustering.

The nearest-neighbor distance distribution, point-event distance distribution and Ripley's K function provide different insights into the spatial pattern. The nn distribution function, $\hat{G}(w)$, is slightly more powerful at detecting departures from CSR in the direction of regularity [24]. The point-event distribution function provides information about the empty space between points. It appears to be more powerful at detecting departures in the direction of clustering [25, 66]. Ripley's K(t) function simultaneously examines the spatial pattern at many distance scales and is now the most popular approach for completely mapped data. However, it is possible to construct point processes with the same $G(w)$ but different Ripley's K(t) function (e.g. [65, p 267]). Conversely, [4] illustrate two processes with the same Ripley's K(t) function but very different nearest-neighbor distance distributions and point-event distance distributions.

Are events points or circles?

All the distributional theory of the previous sections assumes that events occupy no space. Treating events as points assumes that it is possible for two events (e.g. locations of tree trunks or bird nests) to be an infinitesimally small distance from each other. The point assumption is reasonable when the area of the events is small relative to the spacing between the events. The assumption is likely to be appropriate for bird nests (generally small) or tree trunks (generally low density), but not for ant nests (large size relative to the density of nests in an area). If events are incorrectly assumed to be points, the analysis of the spatial pattern indicates a tendency to regularity because two events do not occur within a small distance (the physical size of the event) of each other.

An approximation to the mean event-event distance for non-overlapping circles under CSR is

$$\overline{W} \approx d + \exp(-\rho\pi d^2)[1 - \Phi(\sqrt{2\rho\pi d})]/\sqrt{\rho}, \tag{9}$$

where d and ρ are the diameter of the circles and the number of circles per unit area [63]. Equation 9 assumes low intensity, ρ, and a small and constant diameter, d. The distribution, $G(w)$, of nearest-neighbor distances for non-overlapping circles can be estimated by Monte-Carlo simulation using a sequential inhibition algorithm [24]. The distribution of event-event distances can be complicated when the circles are large or the density is high. For example, it may not be possible to fit the required number of large circles into the study area.

Algorithms and computing

The simplest way to compute nearest neighbor distances is direct enumeration, i.e. computing the distances between all pairs of points, then reporting the smallest distance for each point as the nearest-neighbor distance. For a large number of points, this becomes impractical and more efficient algorithms have been developed. Possible approaches include subdividing the region into smaller subregions [47], computing the Dirichlet tessellation (also known as the Voronoi tessellation, Thiessen tessellation, and other less common names [66, p. 96]) and using that to identify nearest neighbors, or computing the quadtree, a sorted matrix of locations that simplifies the search for nearest neighbors, and using that to identify nearest neighbors [35]. Murtagh [47] reviews the properties of these algorithms.

Functions or procedures for nearest-neighbor spatial analysis are included in few statistical programs, but direct enumeration is very simple to program when needed. Packages of functions for spatial point pattern analysis usually include functions for point-point and point-event analyses. Many of these are S or Splus libraries, e.g. Splancs [60], spatial [67], and S+SpatialStats [45].

Example: trees in a swamp hardwood forest

Figure 1 shows the locations of all 630 trees (stems > 11.5 cm diameter at breast height) and the locations of 91 cypress trees in a 1 ha plot of swamp hardwood forest in South Carolina, USA. There are 13 different tree species in this plot, but over 75% of the stems are one of three species, black gum, *Nyssa sylvatica*, water tupelo *Nyssa aquatica*, or bald cypress, *Taxodium distichum*. Visually, trees seem to be scattered randomly throughout the plot, but cypress trees seem to be clustered in three bands. Nearest-neighbor statistics provide a way to test the hypothesis that stems are randomly distributed throughout the plot. I will illustrate tests based on mean nearest-neighbor distance, G(w), and F(x) using the locations of all 630 trees and the locations of the 91 cypress trees.

For all 630 tree locations, the mean nearest-neighbor distance is 1.99m. If 630 points were randomly distributed in a 200m x 50m rectangle, the expected nn distance is 2.034m, with a s.e. of 0.044m, using Donnelly's approximations (equations 6 and 7). There is no evidence of departure from CSR (z = -0.973 with a 2 sided p-value of 0.33). The effect of the edge corrections is minimal probably because the plot is large and the nearest-neighbor distance is small. The uncorrected expected nn distance is 1.992m, with a s.e. of 0.042m, using equations 3 and 4.

Conclusions using the distribution of nearest-neighbor distances, $\hat{G}(w)$, are similar. $\hat{G}(w)$ was estimated without edge correctionsm, so $\hat{G}(w)$ must be compared to simulated values, not the theoretical expectation (equation 2). The observed cdf, the theoretical expected value (equation 2) and the average simulated cdf are very similar (Figure 2a), although the observed $\hat{G}(w)$ is slightly larger than the expected value at short distances. The differences can be seen more clearly if $1 - \exp(-\rho\pi w^2)$ (equation 2) is subtracted from all curves (Figure 2b). Although $\hat{G}(w)$ is larger than both the theoretical expected value and the average simulated value, it lies within the pointwise 95% confidence limits. None of the three summary statistics (Kolmgorov-Smirnov, Cramer-Von Mises, or Anderson-Darling) is significant at $\alpha = 0.05$. For example, the observed Kolmogorov-Smirnov test statistic of 0.044 is less than the simulated 90th percentile, 0.052. The estimated p-values for the Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling test statistics are 0.19, 0.31, and 0.40.

In contrast, the distribution of point-event distances suggests there is some clustering of tree locations. The observed $\hat{F}(x)$ is below the theoretical and average simulated curves (Figure 3a, b) and outside the pointwise 95% confidence bounds at large distances (Figure 3b). Because $\hat{F}(x)$ falls below the expected values, distances from randomly chosen points are stochastically greater than expected if events were CSR. The greater than expected abundance of large empty spaces provides evidence of clustering of the events. P-values for the Kolmogorov-

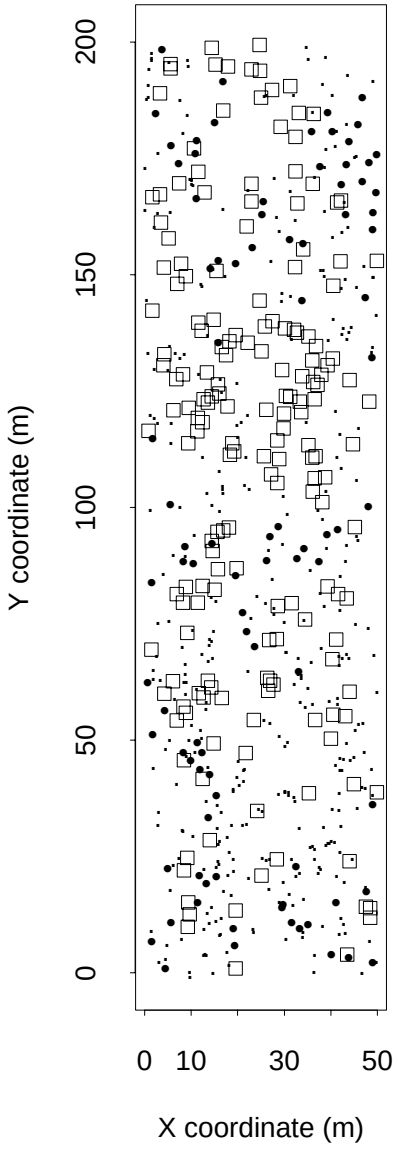

Figure 1: Marked plot of tree locations in a 50m x 200m plot of hardwood swamp in South Carolina, USA. Circles are locations of cypress trees, squares are locations of black gum trees, and dots are locations of any other species.

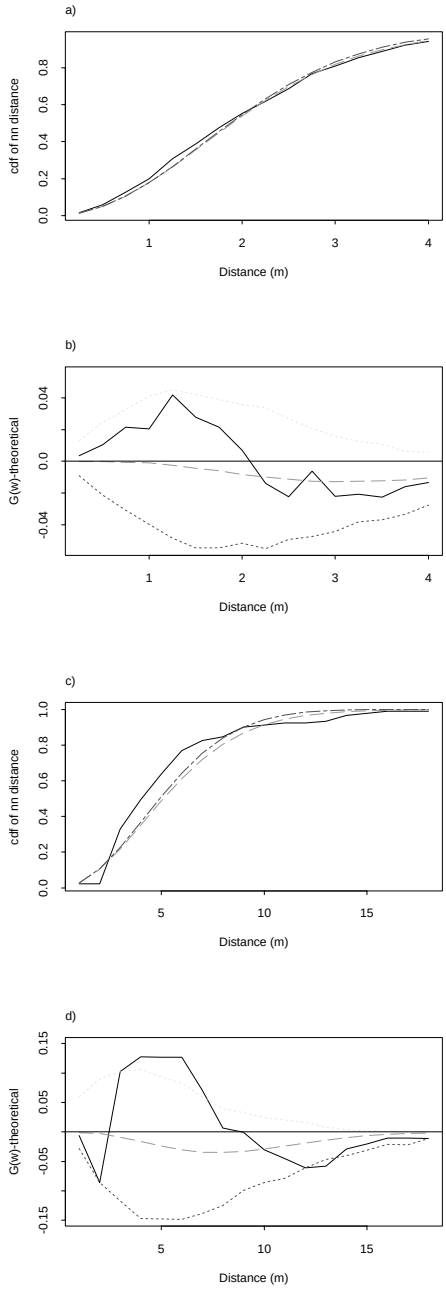

Figure 2: Cdf plots of nearest-neighbor distances for all trees and for cypress only.

a) All trees, comparison of $\hat{G}(w)$ (solid line), theoretical G(w) (dashed line) and average simulated $\hat{G}(w)$ under complete spatial randomness in a 50m x 200m plot (dotted line).

b) All trees, comparison of cdfs. The theoretical G(w) subtracted for clarity. $\hat{G}(w)$ is the solid line, the average simulated $\hat{G}(w)$ is the dotted line, the solid horizontal line is a reference line for the theoretical $G(w)$. The dashed lines are pointwise 0.025 and 0.975 quantiles for $\hat{G}(w)$ under complete spatial randomness.

c) Cypress trees only, comparison of cdfs. Line types are the same as in a).

d) Cypress trees only, comparison of cdfs minus the theoretical G(w). Line types are the same as in b).

Smirnov, Cramer-von Mises and Anderson-Darling summary statistics range from 0.004 to 0.009. The conclusion of some evidence for clustering of all tree locations matches the conclusion using Ripley's K function.

For the 91 cypress trees, the mean nn distance is 5.08 m, which is slightly smaller than the edge-corrected expected distance of 5.55m, with a s.e. of 0.33m. Using the nearest-neighbor distance, there is no evidence of a non-random distribution; the z statistic is -1.41, with a 2-sided p-value of 0.16. The effect of the edge corrections is larger when the density of points is smaller. The uncorrected expected nn distance for the 91 Cypress trees is 5.24, with a s.e. of 0.29.

However, the distributions of $\hat{G}(w)$ and $\hat{F}(x)$ for the 91 cypress trees provide evidence of clustering of cypress trees. There are an unusually large number of neighbor-neighbor distances between 3m and 7m (Figure 2c, 2d). This excess is significant; $\hat{G}(w)$ is at or above the point-wise 0.975 quantiles of simulated values (Figure 2d). This excess is consistent with clustering of cypress trees. There are also significant fewer (at least pointwise) nearest-neighbor distances at 13 m. All three summary statistics are significant (Kolmogorov-Smirnov p-value = 0.034, Cramer-von Mises p-value = 0.007, Anderson-Darling p-value = 0.011). The point-event distances are stochastically greater than expected under CSR (Figure 3c, d). The observed distribution, $\hat{F}(x)$ is lies outside the point-wise 95% confidence bands for many distances. All three summary statistics are highly significant (p = 0.001). The conclusion that cypress trees are strongly clustered matches that using Ripley's K function.

Directed tests

The tests in the previous section are general tests of complete spatial randonmess against an unspecified alternative. Other tests may be more powerful when the alternative is more specific (e.g. events are associated with specific sites or the density of events increases from east to west). Association between point events and a non-point stochastic process can be tested using the "nearest-neighbor" distance from each event to nearest part of the second process [7]. However, most directed tests (e.g [43, 68, 55]) use features other than nearest-neighbor distance.

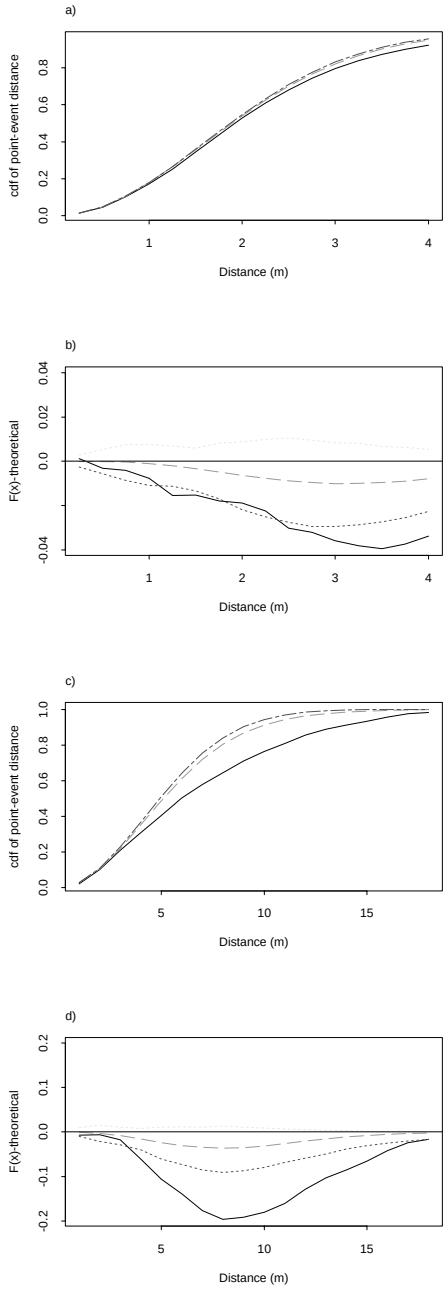

Figure 3: Cdf plots of point-event distances for all trees and for cypress only.

a) All trees, comparison of $\hat{F}(x)$ (solid line), theoretical F(x) (dashed line) and average simulated $\hat{F}(x)$ under complete spatial randomness in a 50m x 200m plot (dotted line).

b) All trees, comparison of cdfs. The theoretical F(x) subtracted for clarity. $\hat{F}(x)$ is the solid line, the average simulated $\hat{F}(x)$ is the dotted line, the solid horizontal line is a reference line for the theoretical $F(x)$. The dashed lines are pointwise 0.025 and 0.975 quantiles for $\hat{F}(x)$ under complete spatial randomness.

c) Cypress trees only, comparison of cdfs. Line types are the same as in a).

d) Cypress trees only, comparison of cdfs minus the theoretical F(x). Line types are the same as in b).

Describing and testing spatial patterns using a sample of nearest neighbor distances.

Although mapped data is often easy to collect, a statistician might view it as wasteful because of the high degree of correlation between nearest neighbor distances (e.g. the perfect correlation for pairs of reflexive nearest neighbors). An alternative is to measure nearest neighbor distances only on a random sample of individuals. Because the nearest-neighbor distances are calculated from a simple random sample of points, the distributional theory for both the mean nn distance and the distribution function is much simpler. Many different tests of CSR have been developed for use with a random sample of point-event, nearest neighbor, or point-event-event distances. These are summarized and evaluated by [66, pp. 59-64] and [18, pp. 602-614].

The most straightforward way to select a random sample of nearest-neighbor distances is to enumerate all individuals in the statistical population, select a simple random sample of events, and measure the distances from the selected events to their nearest neighbors. This can be time consuming and is usually impractical [66]. Enumeration can be avoided by clever use of subregions (described by Byth and Ripley [14]), or by randomly selecting points (not events). The distance from the randomly selected point to the nearest event is a random sample from the distribution of point-event distances, $F(x)$, but the distance from that event to its closest event is not a random sample from $G(w)$ because the point-event and event-event distances are correlated. The distributions of all quantities in the point-event- event sample when events are CSR have been derived [17].

An alternative that is easy to implement in the field is the T-square sample [8], illustrated in [66, 18], a modified point-event-event sample. A point, A, is randomly chosen and the nearest neighbor, B, is found. Then, the study area is divided into two half planes by a line through B and perpendicular to AB (hence the name, T-square). Attention is restricted to the half plane that does not contain point A. The distance to the nearest neighbor, Z, of B in that half plane is measured. When points are CSR, Z and X (the distance from point A to nearest neighbor B) are independent, and the distribution of $Z/\sqrt{2}$ is the same as the distribution of nearest- neighbor distances, $G(w)$ [8].

Estimating density

A random sample of point to nearest neighbor distances can be used to estimate ρ, the average density of events in the study area. When events are CSR, the maximum likelihood estimate is

$$\hat{\rho} = n \left(\pi \sum_{i=1}^{n} X_i^2 \right)^{-1}, \tag{10}$$

where X_i is the distance from a randomly chosen point to its nearest neighbor [46]. An unbiased estimate is [54]

$$\hat{\rho}_P = (n-1) \left(\pi \sum_{i=1}^{n} X_i \right)^{-2}, \tag{11}$$

Both estimators are very dependent on the CSR assumption and can be biased if locations are clustered or regularly distributed.

Many other estimators have been proposed. Upton and Fingleton [66, pp. 118-133] summarize and provide examples of calculations for many of these. Byth [13] evaluated robustness of many estimators to deviations from CSR. She recommended an estimator based on two quantities from T-square sampling, X the distance from point to nearest event and Z, and the distance to nearest-neighbor in a half-plane.

$$\hat{\rho}_T = n^2 \left[2\sqrt{2} \left(\sum_{i=1}^{n} X_i \right) \left(\sum_{i=1}^{n} Z_i \right) \right]^{-1}. \tag{12}$$

Nearest neighbor methods to examine spatial patterns of more than one type of point

Additional information about an event is often available. For example, tree locations might be marked with the species of tree (a mark with discrete levels) or the size of the tree (a mark with continuous levels). The methods in the previous sections can be used to analyze patterns in all events (ignoring the marks) or subgroups of events (e.g. just species A or just trees larger than 50cm). However, other interesting questions could be asked about the relationship between the two (or more) sets of locations.

Multivariate spatial point patterns are those where events can be classified into different types, i.e. the marks are discrete [18, p. 707]. Usually, the number of different types is small; bivariate patterns, with two types of marks, are the most common. Some questions that could be asked about point processes with discrete marks are:

1. Are the processes that generate locations with different marks independent?

2. Are marks randomly assigned to locations? Conditional on the observed locations of superposition of the two marked processes, are the marks independent?

3. Are marks segregated? Are locations with one type of mark surrounded by locations with the same mark?

These questions about the relationships between processes make no assumptions about the marginal pattern of each process. In particular, either process (or their superposition) may be independent, clustered, or regularly distributed. The two general methods to answer these questions are the comparison of distribution functions [38, 44] and the nearest-neighbor contingency table [52, 30]. Other approaches include Ripley's K(t) functions or parametric point process models.

Define the following multi-type extensions of the point-event and nearest-neighbor distances. X_i is the distance from a randomly chosen point to the nearest event with mark i, with cdf $F_i(x)$. W_{ij} is the distance from an event with mark i to the nearest event with mark j, with cdf $G_{ij}(w)$. If the process with mark i is independent of the process with mark j, then:

$$F_i(x) = G_{ji}(x) \tag{13}$$

$$F_j(x) = G_{ij}(x), \tag{14}$$

and X_i and X_j are independent [38, 28]. Note that property (13) does not imply property (14), so two tests are needed [38].

For sparsely sampled data, Goodall [38] suggests a t-test of $\overline{X}_i = \overline{X}_j$. Diggle and Cox [28] consider non-parametric versions of the t-test, tests of equality of distribution, and tests of the correlation between X_i and X_j. Details and a comparison of the tests are given in Diggle and Cox [28]. Analyses of completely mapped data tend to focus on the comparison of distribution functions in equations (13) and (14). Monte-Carlo tests are used because of the non-independence of point-event and event-event distances.

Two different simulation methods could be used in the Monte-Carlo test. The choice depends on the null hypothesis. If the null hypothesis is independence between marks (question 1, above), then toroidal shifts or some parametric model should be used to generate the randomization distribution. If the null hypothesis is random assignment of marks conditional on the set of events, then random labelling of events should be used to generate the randomization distribution. In general, these two hypotheses are not equivalent and the sampling distributions are not the same.

Nearest-neighbor contingency tables

The nearest-neighbor contingency table focuses on the ecologically important question of segregation [52, 30]. This table describes marks of events and their nearest neighbors, not the distance between them (Table 1). In sparsely sampled data, the counts (N_{AA}, N_{AB}, N_{BA}, and N_{BB}) are independent Poisson random variables or conditionally independent given the row marginal totals (N_A and N_B) under the null hypothesis of random labelling [52]. The hypothesis can be tested with a traditional 1 df Chi-square test of independence [52].

	Mark of neighbor		
Mark of point	A	B	Total
A	N_{AA}	N_{AB}	N_A
B	N_{BA}	N_{BB}	N_B
Total	M_A	M_B	N

Table 1: Cell counts in nearest-neighbor contingency table

In completely mapped data, the sampling distribution of the counts is different. [16]. If events are randomly labelled, the expected values of the counts depend only on the number of each type of event (N_A and N_B) and the total number of events, N [30] (Table 2). The variances and covariances depend on the number of events of each type, the number of reflexive nearest neighbors, and the number of shared nearest-neighbors; Dixon [30] derives the formulae. The first two moments of the cell counts can be used to test for segregation of type A events ($N_{AA} >$ E N_{AA}), test for segregation of type B events ($N_{BB} >$ E N_{BB}), or construct an omnibus 2 d.f. Chi-square test of random labelling. If the numbers of points are large, the distributions of test statistics can be adequately approximated by the asymptotic normal and Chi-square distributions. If the number of points is small, the distributions should be determined by Monte-Carlo simulation.

	To: A	To: B
From: A	$N_A(N_A-1)/(N-1)$	$N_AN_B/(N-1)$
B	$N_BN_A/(N-1)$	$N_B(N_B-1)/(N-1)$

Table 2: Expected cell counts for nearest-neighbor contingency table

Patterns with k marks ($k > 2$) can be analyzed by considering all pairs of marks two at a time (using distance methods or 2 x 2 nearest-neighbor contingency tables), or by considering the k x k contingency table. The expected counts and their variances under random labelling follow the same form as those for a 2 x 2 contingency

table, but there are more possible forms for the covariance between two counts. Details are given in [29].

Other approaches that have been suggested for the analysis of multi-type point processes include the comparison of bivariate Ripley's K functions [44, 27], empty space methods [44] (comparisons of point-event distance distributions), and mark correlation functions [65].

Example, part 2

Cypress and black gum are two of the three abundant species in the 1 ha plot of swamp forest considered previously. An interesting ecological question is whether those two species are spatially segregated, that is do cypress trees tend to be found near other cypress trees and do black gum trees tend to be found near other black gum trees? The marked plot of locations (Figure 1) suggests that cypress trees and black gum occur in different clusters. Confirming this requires an analysis of the bivariate spatial pattern. The three tests that will be illustrated are the comparisons of cdfs (equations 13 and 14)[28], the independence of distances [28] and the nearest-neighbor contingency table (Figure 1) [30]. The ecological background suggests that random labelling is the more appropriate null hypothesis.

The cdf of distances from randomly chosen points to the nearest black gum, $F_G(x)$, and the cdf of point-event distances to the nearest cypress, $F_C(x)$, were estimated without edge-corrections using a randomly located grid of points [14]. The cdf's of distances from black gums to the nearest cypress, $G_{GC}(x)$, and from cypresses to the nearest black gum, $G_{CG}(x)$, were also estimated without edge corrections. Both species show the same pattern. Cypress trees are stochastically further from black gum trees than from randomly chosen points (Figure 4a). Also, black gum trees are stochastically further from cypresses than from randomly chosen points (Figure 4b).

The observed differences can be compared to those found under random labelling by using the Kolmogorov-Smirnov 2 sample statistics, max $\mid F_G(x) - G_{CG}(x) \mid$ and max $\mid F_C(x) - G_{GC}(x) \mid$, as the test statistics. The observed values differences are not unusual large (p-value = 0.109 for black gum and 0.083 for cypress). The distance from a randomly chosen points to the nearest black gum, X_G, is negatively correlated with the distance from the same point to the nearest cypress, X_C (Kendall's $\hat{\tau}$ = -0.12, with a one-sided p-value by randomization of 0.001). This result is consistent with spatial segregation of the two species. The different results from the three tests are consistent with Diggle and Cox's [28] observation that the correlation test is more powerful than the Kolmogorov-Smirnov test for the sparsely sampled spatial patterns they studied.

The nearest-neighbor contingency table indicates that both species have an excess of nearest neighbors of the

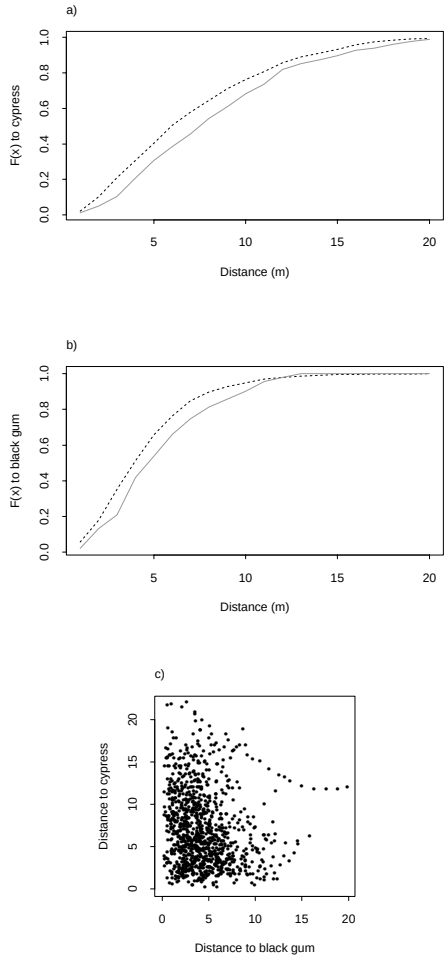

Figure 4: Cdf plots of point-event distances for all trees and for cypress only.

a) Comparison of cdfs of point to cypress distances ($F_C(x)$, dotted line) and black gum to cypress distances ($G_{GC}(x)$, solid line).

b) Comparison of cdfs of point to black gum distances ($F_G(x)$, dotted line) and cypress to black gum distances ($F_{CG}(x)$, solid line).

c) Relationship between distances from a randomly chosen point to the nearest cypress and nearest black gum.

same species (table 3). The variances of the cell counts are 38.88 for black gum - black gum and 25.55 for cypress-cypress. P-values can be computed by Monte-Carlo randomization or by a normal approximation [30]. Either way, the one-sided p-values are small (0.001 or < 0.001) for both species.

	Species of neighbor		
Species of point	Black gum	Cypress	Total
Black gum	149 (121.1)	33 (60.9)	182
	4.47	**-4.47**	
Cypress	43 (60.9)	48 (30.1)	91
	-3.54	**3.54**	
Total			273

Table 3: Observed counts, (expected counts), and z scores (in bold) for the cypress and black gum nearest-neighbor contingency table.

Nearest Neighbor Methods for Field Experiments

A different set of nearest-neighbor methods can be used to analyze spatially structured field experiments. 0ne of the most common applications is agronomic variety trials, where many treatments are compared using small plots arranged in a rectangular lattice. Traditional methods of controlling for between-plot heterogeneity, such as using a randomized complete block design, may not be very effective because the large number of treatments forces the blocks to be large. Nearest-neighbor methods use information from adjacent plots to adjust for within-block heterogeneity and so provide more precise estimates of treatment means and differences. If there is within-plot heterogeneity on a spatial scale that is larger than a single plot and smaller than the entire block, then yields from adjacent plots will be positively correlated. Information from neighboring plots can be used to reduce or remove the unwanted effect of the spatial heterogeneity, and hence improve the estimate of the treatment effect. Data from neighboring plots can also be used to reduce the influence of competition between adjacent plots. Each of these approaches will be briefly discussed.

Papadakis [49, 50] proposed an analysis of covariance to reduce the effects of small-scale spatial heterogeneity in yields. The value of the covariate for each plot is obtained by averaging residuals from the neighboring plots. The choice of neighboring plots depends on the crop, the plot size and shape, and the spacing between plots. In many row crops, the neighboring plots are defined as the two adjacent plots in the row, except that

plots at an end of a row have only one neighbor. In other situations, it may be appropriate to consider four neighbors, which include the two between-row neighbors. If the spatial heterogeneity is such that the effects of in- and between- row neighbors are quite different, then one could compute separate covariates for within- and between-row neighbors. Once the covariates are computed, treatment effects are re-estimated using an analysis of covariance. For adjustment in one dimension (e.g. along crop rows), the model would be:

$$Y_i = \mu + X_i\tau_i + \beta R_i + \varepsilon_i, \tag{15}$$

where Y_i is the observed yield on the i'th plot, R_i is the mean residual on neighbors of the i'th plot, β is the spatial dependence parameter, μ and τ_i are the parameters in the model for the treatment effects, X_i is the row of the design matrix for the i'th plot and ε_i are the residual variabilities in yields, which are assumed to be uncorrelated. When β is 0, observations on adjacent plots are independent; larger positive values of β ($\beta < 1$) correspond to increasing spatial correlation between neighboring plots. The observed value of β depends on plot size, plot shape, plot spacing, and the scale of the spatial heterogeneity. Values are often close to 1 when plots are small. In the absence of treatments and ignoring edge effects, the Papadakis model implies that correlations between plot yields has a first order autoregressive structure, with $\mathrm{Corr}(Y_i, Y_j) = \lambda^{|i-j|}$, where $\beta = 2\lambda/(1+\lambda^2)$. Values of β close to 1 imply that λ is also close to 1.

The autoregressive correlation structure implied by the ad hoc Papadakis model is one example of the random field approach to a spatially designed experiment [73, 18]. Many other models, including the iterated Papadakis method, the Wilkinson NN [71] model, the Besag and Kempton [9] first order difference models, The Williams model [69] and the Gleeson-Cullis ARIMA models [37, 19] correspond to different specifications of a spatial correlation matrix. Computations are handled either by a general purpose REML algorithm for linear mixed models (e.g. PROC MIXED in SAS, lme() in Splus), or by specialized software for a particular model (e.g. TwoD for 2 dimensional Gleeson-Cullis models [36].

The properties of these methods have been extensively discussed over the last twenty years. Dagnelie [21, 22] provides a relatively recent review and historical summary of the Papadakis model. Wu and Dutilleul [72] use uniformity trial data to compare autoregressive models, difference models, and traditional RCB analyses. The efficiency of a spatial analysis, relative to a randomized complete block design, is usually greater than 1.2 and can be as high as 2 [22]. However, it can give biased estimates of treatment effects [71]. The Papadakis method appears to work best when there are at least three replicates per treatment, many treatments (greater than 10), and strong, but patchy spatial heterogeneity [22]. When there is an underlying trend, first order difference models appear to work well.

Medium-scale spatial heterogeneity usually causes a postive correlation between adjacent plots. When there is competition between plots, neighbors can have a negative effect on the response in adjacent plots [41]. The Papadakis model (15) can be extended to estimate treatment-specific competitive effects. The choice of covariate should be influenced by biological mechanisms. If competition for light is important, a reasonable covariate could be the difference between the mean height of plants in the plot and the mean height on neighboring plots. If disease spread is important, a reasonable covariate could be the mean disease severity on neighboring plots [42]. The coefficient for the covariate (β in equation 15) estimates the strength of the competitive relationship.

Experimental design for a study that will use some form of neighbor adjusted analysis usually focuses on neighbor balance. That is, ensuring that all pairs of treatments occur in adjacent plots equally frequently. Adjacent plots can be defined as only those within the same row (1 dimensional neighbor balanced designs [70]) or as including both those in the same row and those in the same column (2 dimensional neighbor balanced designs [34]). The choice will depend on the size, shape and spacing of the plots and on the biological and physical mechanisms influencing the correlation between plots. Methods for construction of 1D or 2D designs can be found in [70, 34, 33].

References

[1] Ashby, E. 1936, Statistical Ecology. Botanical Review, 2, 221-235.

[2] Ashby, E. 1948, Statistical Ecology. II - A reassessment. Botanical Review, 14, 222-234.

[3] Baddeley, A. and Gill, R. D. 1997, Kaplan-Meier estimators of distance distributions for spatial point processes, Annals of Statistics, 25, 263-292.

[4] Baddeley, A. J. and Silverman, B. W. 1984, A cautionary example on the use of second-order methods for analyzing point patterns. *Biometrics* 40, 1089-1093.

[5] Barnard, G.A. 1963. Contribution to the discussion of Prof. Bartlett's paper. Journal of the Royal Statistical Society, Series B, 25, 294.

[6] Bartlett, M. S. 1978. Nearest neighbour models in the analysis of field experiments (with discussion). Journal of the Royal Statistical Society, Series B, 40, 147-174.

[7] Berman, M. 1986. Testing for spatial association between a point process and another stochastic process. Applied Statistics 35, 54-62.

[8] Besag, J. E. and Gleaves, J. T. 1973. On the detection of spatial pattern in plant communities. Bulletin of the International Statistical Institute, 45, 153-158.

[9] Besag, J. and Kempton, R. 1986. Statistical analysis of field experiments using neighbouring plots. Biometrics, 42, 231-251.

[10] Brewer, R. and McCann, M. T. 1985. Spacing in acorn woodpeckers. Ecology, 66, 307-308.

[11] Brown, 1975. A test of randomness in nest spacing. Wildlife, 26, 102-103.

[12] Brown, D. and Rothery, P. 1978. Randomness and local regularity of points in a plane. Biometrika 65, 115-122.

[13] Byth, K. 1982. On robust distance-based intensity estimators. Biometrics, 38, 127-135.

[14] Byth, K. and Ripley, B. D. 1980. On sampling spatial patterns by distance methods. Biometrics, 36, 279-284.

[15] Clark, P. J. and Evans, F. C. 1954, Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology, 35, 445-453.

[16] Cox, T. F. 1981, Reflexive nearest neighbours. Biometrics, 37, 367-369.

[17] Cox, T. F. and Lewis, T. 1976, A conditioned distance ratio method for analyzing spatial patterns. Biometrika, 63, 483-491.

[18] Cressie, N. 1991. *Statistics for Spatial Data*, Wiley, New York.

[19] Cullis, B. R. and Gleeson, A. C. 1991. Spatial analysis of field experiments - extension to two dimensions. Biometrics, 47,1449-1460.

[20] Cuzick, J. and Edwards, R. 1990. Spatial clustering for inhomogeneous populations (with discussion). Journal of the Royal Statistical Society, Series B, 52, 73-104.

[21] Dagnelie, P. 1987. La méthode de Papadadis en expérentation agronomique: considérations historiques et bibliographiques. Biometrie et Praximetrie, 27, 49-64.

[22] Dagnelie, P. 1989, The method of Papadakis in agricultural estimation: an overview. Biuletyn Oceny Odmian, 21, 111-122.

[23] Devroye, L. and Györfi, L. 1985, *Nonparametric Density Estimation. The L_1 view.* Wiley, New York.

[24] Diggle, P. J. 1979. On parameter estimation and goodness-of-fit testing for spatial point patterns. Biometrics, 35,87-101.

[25] Diggle, P. J. 1983. *Statistical Analysis of Spatial Point Patterns*, Academic Press, London.

[26] Diggle, P. J., Besag, J., and Gleaves, J. T. 1976, Statistical analysis of spatial point patterns by means of distance methods. Biometrics 32, 659-667.

[27] Diggle, P. J. and Chetwynd, A. G. 1991. Second-order analysis of spatial clustering for inhomogeneous populations. Biometrics, 47, 1155-1163.

[28] Diggle, P. J. and Cox, T. F. 1983. Some distance-based tests of independence for sparsely-sampled multivariate spatial point patterns. International Statistical Review, 51, 11-23.

[29] Dixon, P. M. 1992. Testing spatial independence in multivariate point processes. Institute of Statistics Mimeo Series #2215, Raleigh NC.

[30] Dixon, P. M. 1994. Testing spatial segregation using a nearest-neighbor contingency table. Ecology, 75, 1940-1948.

[31] Donnelly, K. P. 1978. Simulations to determine the variance and edge effect of total nearest-neighbour distance. pp 91-95 in Hodder, I. (ed.) *Simulation Studies in Archaeology.* Cambridge University Press, Cambridge.

[32] Everitt, B. S. 1993, *Cluster Analysis, 3rd. edition*, Arnold, London.

[33] Federer, W. T. and Basford, K.E. 1991, Competing effects designs and models for two-dimensional field arrangements. Biometrics, 47, 1461-1472.

[34] Freeman, G. H. 1979. Some two-dimensional designs balanced for nearest neighbours. Journal of the Royal Statistical Society, Series B, 41, 88-95.

[35] Friedman, J., Bentley, J.L. and Finkel, R.A. 1977, An algorithm for finding best matches in logarithmic expected time. ACM Transactions on Mathematical Software 3, 209-226.

[36] Gilmour, A. R. 1992, TwoD. A program to fit a mixed linear model with two dimensional spatial adjustment for local trend. NSW Agriculture Research Center, Tamworth, Australia.

[37] Gleeson, A. C., and Cullis, B. R. 1987. Residual maximum likelihood (REML) estimation of a neighbour model for field experiments. Biometrics 43, 277-288.

[38] Goodall, D.W. 1965. Plot-less tests of interspecific association, Journal of Ecology, 53, 197-210.

[39] Hodder, I. and Orton, C. 1976, *Spatial Analysis in Archaeology.* Cambridge Univ. Press, Cambridge.

[40] Gignoux, J., Duby, C. and Barot, S. 1999. Comparing the performance of Diggle's tests of spatial randomness for small samples with and without edge-effect correction: application to ecological data. Biometrics, 55, 156-164.

[41] Kempton, R. 1982, Adjustment for competition between varieties in plant breeding trials. Journal of Agricultural Science 98, 599-611.

[42] Kempton, R. 1991, Interference in agricultural experiments. Proceedings of the Second Biometric Society East / Central / Southern African Network Meeting, Harare, Zimbabwe.

[43] Lawson, A. 1988, On tests for spatial trend in a non-homogeneous Poisson process. Journal of Applied Statistics 15, 225-234.

[44] Lotwick, H.W. and Silverman, B. W. 1982. Methods for analyzing spatial processes of several types of points. Journal of the Royal Statistical Society, Series B. 44, 406-413.

[45] Mathsoft 1996, *S+SPATIALSTATS User's Manual, Version 1.0*, Mathsoft Inc., Seattle.

[46] Moore, P. G. 1954, Spacing in plant populations, Ecology, 35,222-227.

[47] Murtagh, F. 1984, A review of fast techniques for nearest neighbor searching, Compstat 1984, 143-147.

[48] Neyman, J. and Scott, E. L. 1952, A theory of the spatial distribution of galaxies, Astrophysical Journal, 116, 144-163.

[49] Papadakis, J. S. 1973, Methode statistique pour des experiences sur champ. Institut d'Amelioration des Plantes a Thessaloniki (Grece) Bulletin Scientifique, No. 23.

[50] Papadakis, J. S. 1984, Advances in the analysis of field experiments. Proceedings of the Academy of Athens, 59, 326-342.

[51] Pickard, D. K. 1982, Isolated nearest neighbors, Journal of Applied Probability 19, 444-449.

[52] Pielou, E. C. 1961. Segregation and symmetry in two-species populations as studied by nearest-neighbour relationships. Journal of Ecology, 49, 255-269.

[53] Pielou, E. C. 1977. *Mathematical Ecology.* Wiley, New York.

[54] Pollard, J.H. 1971, On distance estimators of density in randomly distributed forestes. Biometrics, 27, 991-1002.

[55] Rathbun, S. L. 1996. Estimation of Poisson intensity using partially observed concomitant variables. *Biometrics* 52, 226-242.

[56] Ripley, B. D. 1977. Modelling spatial patterns (with discussion). Journal of the Royal Statistical Society, Series B, 39, 172-212.

[57] Ripley, B. D. 1979. Tests of 'randomness' for spatial point patterns. Journal of the Royal Statistical Society, Series B. 41, 368-374.

[58] Ripley, B. D. 1981, *Spatial Statistics*. Wiley, New York.

[59] Ripley, B. D. 1996, *Pattern Recognition and Neural Networks*. Cambridge Univ. Press., Cambridge

[60] Rowlinson, B.S. and Diggle, P. J. 1992, Splancs: Spatial Point Pattern Analysis Code in S-Plus. Technical Report 92/63, Lancaster University, U.K.

[61] Schilling, M. F. 1986, Mutual and shared neighbor probabilities: finite- and infinite-dimensional results. Advances in Applied Probability 18, 388-405.

[62] Silverman, B. and Brown, T. 1978, Short distance, flat triangles, and Poisson limits. Journal of Applied Probability, 15, 815-825.

[63] Simberloff, D. 1979, Nearest neighbor assessments of spatial configurations of circles rather than points. Ecology, 60, 679-685.

[64] Stoyan, D. and Penttinen, A. 2000, Recent applications of point process methods in forestry statistics. Statistical Science 15, 61-78.

[65] Stoyan, D. and Stoyan, H. 1994, *Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics*. Wiley, Chichester.

[66] Upton, G. J. G. and Fingleton, B. 1985. *Spatial Data Analysis by Example, Volume 1. Point pattern and quantitative data*. Wiley, Chichester.

[67] Venables, W. N. and Ripley, B. D. 1994, *Modern Applied Statistics with S-Plus*, Springer-Verlag, New York.

[68] Waller, L. A., Turnbull, B. W., Clark, L. C. and Nasca, P. 1994, Spatial pattern analyses to detect rare disease clusters. pp 3-23 in Lange, N. et al. (eds). *Case Studies in Biometry*, Wiley, New York.

[69] Williams, E. R. 1986, A neighbour model for field experiments, Biometrika, 73, 279-287.

[70] Williams, R. M. 1952, Experimental designs for serially correlated observations. Biometrika 39, 151-167.

[71] Wilkinson, G. N., Eckert, S. R., Hancock, T. W., and Mayo, O. 1983. Nearest neighbour (NN) analysis of field experiments (with discussion). Journal of the Royal Statistical Society, Series B, 45, 151-211.

[72] Wu, T. and Dutilleul, P. 1999. Validity and efficiency of neighbor analyses in comparison with classical complete and incomplete block analyses of field experiments. Agronomy Journal 91:721-731.

[73] Zimmerman, D. L. and Harville, D. A. 1991, A random field approach to the analysis of field-plot experiments and other spatial experiments. Biometrics, 47, 223-239.